npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Check for updates

Towards the automatic detection of social biomarkers in autism spectrum disorder: introducing the simulated interaction task (SIT)

Hanna Drimalla[1,2,3]✉, Tobias Scheffer[4], Niels Landwehr[4,5], Irina Baskow[1,6], Stefan Roepke[6], Behnoush Behnia[6,7] and Isabel Dziobek[1,2,7]

Social interaction deficits are evident in many psychiatric conditions and specifically in autism spectrum disorder (ASD), but hard to assess objectively. We present a digital tool to automatically quantify biomarkers of social interaction deficits: the simulated interaction task (SIT), which entails a standardized 7-min simulated dialog via video and the automated analysis of facial expressions, gaze behavior, and voice characteristics. In a study with 37 adults with ASD without intellectual disability and 43 healthy controls, we show the potential of the tool as a diagnostic instrument and for better description of ASD-associated social phenotypes. Using machine-learning tools, we detected individuals with ASD with an accuracy of 73%, sensitivity of 67%, and specificity of 79%, based on their facial expressions and vocal characteristics alone. Especially reduced social smiling and facial mimicry as well as a higher voice fundamental frequency and harmony-to-noise-ratio were characteristic for individuals with ASD. The time-effective and cost-effective computer-based analysis outperformed a majority vote and performed equal to clinical expert ratings.

npj Digital Medicine (2020)3:25 ; https://doi.org/10.1038/s41746-020-0227-5

INTRODUCTION

Social interaction deficits, which encompass non-verbal communication behavior, such as reading and signaling emotions, are prevalent in many psychiatric disorders.[1] Difficulties in the encoding of emotions are a defining characteristic of autism spectrum disorder (ASD),[2,3] but have also been reported for other psychiatric disorders.[4–7] Similarly, differences in signaling emotions are characteristic for ASD[8] and common in many other mental disorders.[9]

When diagnosing and monitoring the course of a mental disorder, it is thus crucial to assess a patient's respective deficits with reliable and time-efficient methods. For an objective description and diagnosis of patients' deficits in general cognition, standardized tests are available, e.g., the Montreal Cognitive Assessment[10] or the Mini Mental State Examination.[11] In contrast, quantifying social interaction deficits is still in its infancy. Apart from a few exceptions, e.g., Magdeburg Test of Social Intelligence[12], clinicians currently rely on the so-called "clinical gaze", the implicit knowledge based on experience, which is demanded to be used in some clinical investigations such as the autism diagnostic observation schedule (ADOS).[13] Practitioners need many years of training to acquire the necessary expertise, which is also difficult to verbalize, teach, quantify, standardize, and validate.

The need for precise and standardized tools to measure social interaction deficits is especially evident in ASD, as deficits in social communication and interaction are the core symptomatology in ASD.[14] The Autism Diagnostic Interview-Revised (ADI-R) and ADOS represent the gold standard of ASD diagnostics and have proven objective and reliable.[13,15] Such clinical measures include evaluating the patient's facial expressivity and gaze behavior by a trained clinician following a standardized protocol. However, interrater-variability may account for inconsistencies regarding diagnostic accuracy.[16] The lack of feasible standardized diagnostic instruments contributes to a high number of non- or late diagnosed individuals.[17] Especially individuals with ASD and average or above average intelligence are often diagnosed later in life,[18] as they develop strategies to compensate for their deficits, which has been referred to as camouflaging.[19] Thus, ASD diagnostics would greatly benefit from automatic methods measuring social interaction deficits validly and reliably.

Digital standardized tests that automatically analyze a patient's social behavior would provide a widely accessible time-efficient and cost-efficient alternative to expert diagnosis. Some studies have recently shown the potential of analyzing social behavior, such as speech,[20] facial expressions,[21] and gaze behavior[22] with machine-learning methods to assess general mental disorder status. However, to date paradigms to measure social behavior in a standardized way, in which an interaction partner reacts in a reproducible and comparable way to each participant, are lacking. Consequently, we designed the simulated interaction test (SIT), entailing a short standardized simulated dialog between a participant and an actress about food preferences and dinner preparation.

The SIT evokes aspects of social interaction behavior previously described as atypical for individuals with ASD: Individuals with ASD share emotions of others less intensely, which has been reported for negative[23] and positive emotions.[24] The mimicry of

[1]Department of Psychology, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany. [2]Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany. [3]Digital Health Center, Hasso Plattner Institute, University of Potsdam, Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, Germany. [4]Institute of Computer Science, University of Potsdam, Am Neuen Palais 10, 14469 Potsdam, Germany. [5]Leibniz Institute for Agricultural Engineering and Bioeconomy, Max-Eyth-Allee 100, 14469 Potsdam, Germany. [6]Department of Psychiatry, Charité-Universitätsmedizin Berlin, Campus Benjamin Franklin, Hindenburgdamm 30, 12203 Berlin, Germany. [7]These authors contributed equally: Behnoush Behnia, Isabel Dziobek. ✉email: hanna.drimalla@hu-berlin.de

facial expressions has been found reduced,[25] an effect that scales with severity of social dysfunction in ASD.[26] Atypical gaze patterns[27,28] are furthermore characteristic of ASD—especially the avoidance of direct eye contact.[29–31] Moreover, ASD involves aberrant voice intonation,[32] especially in naturalistic settings.[33] The SIT aims at reliably capturing those social biomarkers, which may aid earlier diagnosis as well as monitoring of the course of the disorder and treatment outcome.

Recent attempts towards meaningful diagnostic predictions have shown the potential of machine-learning approaches to analyze non-verbal behavior differences to detect ASD.[34–38] However, to the best of our knowledge, no existing approach focuses on adults with ASD and normal intelligence levels despite the high need for access to diagnosis.[39,40] Thus, using computer vision and machine learning on video- and audio recordings, in the current study we aimed to identify social biomarkers of ASD that allow an affordable, accessible, and time-effective identification of the diagnosis.

This paper encompasses two studies: In a facial electromyography (EMG) preparatory study, we assessed facial behavior via the SIT in a sample of healthy controls (HC; i.e., individuals without autism). Making use of the high precision of EMG, we aimed at the precise description of non-verbal facial behavior in the simulated interaction to select relevant features (i.e., regions of interest) for the main study. In the ASD study, we aimed to replicate the results for HC using automated methods and compared non-verbal social communication behavior of individuals with and without ASD. Further, we followed up on previous research[41] predicting the diagnosis of ASD individuals based on interaction behavior, comparatively evaluating the SIT's diagnostic properties and gold standard clinical measures as well as judgments of clinical experts.

RESULTS

Participants

We analyzed the video- and audio-recordings of 120 participants in total, which took the SIT, 80 (ASD: 37, NT: 43) in the clinical main study and 40 male HC in the EMG preparatory study (see Supplementary Information). In the main study, all participants' video recordings were analyzed by clinical experts as well as by computer-based tools.

Computer-based analysis

The results of the computer-based analysis showed that individuals with ASD could be detected with varying accuracy based on facial expressions, gaze behavior, or vocal characteristics. For transparent and explainable diagnostic decisions, we limited ourselves strictly to features based on domain-knowledge or on the EMG preparatory study (see Supplementary Information). Due to the strong skewness and non-normality of the data, we only used non-parametric tests.

Nearly all frames (99% in both groups) of the video recording data could be analyzed successfully with the computer vision tool OpenFace[42] with high confidence (on a scale of 0–1: mean = 0.99 in both groups, std = 0.013). There was no evidence that one of the two groups (ASD, NT) could be tracked with higher confidence ($p = 0.97$). We excluded all frames that were either not tracked successfully or with confidence below 0.75. For a fair comparison of the classifiers, we had to exclude one female participant with autism because of missing audio recording.

First, we investigated facial expressions across groups and examined whether participants expressed the relevant action unit (AU) for the part of the communication involving positive emotions (AU12 and AU6) and negative emotions (AU4). To account for participants' baseline facial expression we compared mean occurrence and intensity of the AUs relevant for each part to the activity during the neutral part of the communication.

In line with the results of the EMG study, participants showed more positive facial expressions during the positive part of the conversation (talking and listening about favorite dishes) compared to the neural/negative parts. The positive expressions were evident in a higher mean intensity of AU12 (Mdn = 0.30) as well as of AU6 (Mdn = 0.43) in comparison to their intensity in the neutral part, AU12 (Mdn = 0.24; $Z = 703$, $p < 0.001$) and AU6 (Mdn = 0.33), $Z = 567$, $p < 0.001$). Likewise, the participants showed higher occurrence of AU12 (positive: Mdn = 0.20 vs. neutral: Mdn = 0.10; $Z = 499.0$, $p < 0.001$) and AU6 (positive: Mdn = 0.04 vs. neutral: Mdn = 0.001; $Z = 232.0$, $p < 0.001$) compared to the neutral part.

In line with the preparatory study, participants tended to show more negative expressions during the negative part of the conversation (talking about disliked food), which was evident as a trend towards a higher intensity of AU4 (Mdn = 0.058) compared to the neutral part of the conversation (Mdn = 0.056), $Z = 1044.0$, $p = 0.06$.

To compare the facial expressions of both groups, we focused first on the two emotional parts of the conversation and the respective AUs. In the positive part of the conversation (favorite dish), individuals with ASD (Mdn = 0.12) expressed AU12 (smiling mouth) compared to HC (Mdn = 0.40) significantly less frequently ($U = 578.0$, $p = 0.018$, $z = 0.27$). Regarding AU6 (smiling eyes), we found no evidence for group differences in intensity or occurrence (all $p > 0.05$) in the positive part of the conversation.

In the negative part, individuals with ASD showed more negative facial expressions than individuals without ASD, i.e., more intense AU4 activity (ASD: Mdn = 0.08 vs. NT: Mdn = 0.03), $U = 568$, $p = 0.014$, $r = 0.29$.

Next, we compared social smiling in both groups during the entire conversation. For both relevant AUs (AU12 and AU6), we inspected occurrence and intensity. Individuals with ASD showed less social smiling in the mouth region (AU12) compared to individuals without ASD; this was evident in AUs' occurrence (ASD: Mdn = 0.09 vs. NT: Mdn = 0.40; $U = 557.0$, $p = 0.01$, $r = 0.30$) as well as intensity (ASD: Mdn = 0.16 vs. NT: Mdn = 0.52; $U = 581.0$, $p = 0.019$, $r = 0.27$). Additionally, individuals with ASD showed less social smiling around their eyes (AU6) compared to individuals without ASD, again evident in occurrence (ASD: Mdn = 0.02 vs. NT: Mdn = 0.12; $U = 598.5$, $p = 0.028$, $r = 0.25$).

To measure facial mimicking behavior of the participants, we calculated the time-shifted-correlation of the actress' facial behavior and the participants' facial expression behavior. Mimicking the actress' AU6 activity was more pronounced in HC. This was evident in a higher correlation of the actress' and participants' time series in this AU's activity in the HC group (ASD: Mdn = 0.08 vs. NT: Mdn = 0.19, $U = 113.0$, $p = 0.04$, $r = 0.34$).

Applying a random forest classifier and all of the facial expression features, we could predict an individual's diagnosis with an area under the ROC curve of AUC = 0.65.[43] Using all 17 AU provided by OpenFace, we were able to predict the diagnosis with AUC of 0.74 AUC. An exploratory in-depth analysis revealed that more female individuals (34 out of 39) than male individuals (20 out of 40) could be classified correctly based on facial expressions ($X = 12.34$, $p < 0.001$).

Figure 1 shows the predictions of the face-based-classifier separated by gender and in comparison to experts' classifications. Adding age as a feature did not improve the accuracy of the classifier.

For gaze analysis, again, we first excluded frames that were tracked non-successfully or with low confidence. Next, we compared the two groups based on gaze variation. We used the individual's gaze angle in radians in world coordinates, averaged for both eyes, as they are provided by OpenFace.[42] Three example participants' raw gaze patterns are displayed in Supplementary Fig. 2.

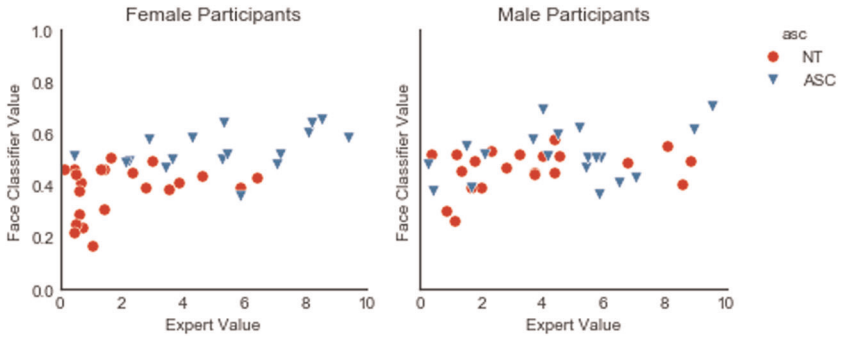

Fig. 1 Automated classification based on facial expression separated by gender.

Table 1. Gaze behavior separated by groups.

	NT	ASD
Mean (gaze angle horizontal)	0.02 (SD: 0.06)	0.01 (SD: 0.07)
Mean (gaze angle vertical)	−0.25 (SD: 0.12)	−0.23 (SD: 0.11)
Absolute deviation from median gaze angle (horizontal)	0.02 (SD: 0.01)	0.03 (SD: 0.01)
Absolute deviation from median gaze angle (vertical)	0.03 (SD: 0.01)	0.033 (SD: 0.01)
Mean speed of eye movement (horizontal)	0.65 (SD: 0.20)	0.67 (SD: 0.21)
Mean speed of eye movement (vertical)	0.71 (SD: 0.32)	0.64 (SD: 0.26)
Mean acceleration of eye movement (horizontal)	0.97 (SD: 0.32)	0.99 (SD: 0.32)
Mean acceleration of eye movement (vertical)	0.93 (SD: 0.42)	0.88 (SD: 0.38)

Table 2. Voice characteristics including harmony-noise-ratio (HNR) and fundamental frequency (F_0) in Hertz (Hz) for both groups.

	NT	ASD
Mean F_0 (Hz) female	209.08 (SD: 17.97)	218.91 (SD: 16.16)
Mean F_0 (Hz) male	121.68 (SD: 11.68)	139.67 (SD: 16.22)
Median HNR female	10.32 (SD: 1.47)	6.91 (SD:2.12)
Median HNR male	10.84 (SD:1.42)	8.50 (SD:1.44)

We compared the groups regarding the following features: difference in variance and means on eye movements of the vertical and the horizontal axis as well as differences in speed of the eye movements. To account for difference in participants' height or position, we also calculated the absolute deviation of the values from the median of the eye gaze direction. In general, the variation around the median gaze angle was small. Correcting for multiple comparisons, we found no evidence for group differences in gaze behavior. The descriptive values for both groups are presented in Table 1.

Using a random forest classifier and all gaze behavior features, we could predict an individual's diagnosis with an AUC = 0.63. The exploratory in-depth analysis of the classification results showed no significant evidence for different accuracy of classifying male and female participants.

The descriptive values of the voice characteristics for both groups are presented in Table 2. As pitch and harmony-noise-ratio (HNR) vary strongly between males and females,[44,45] we included participants' gender in the analysis. A significant main effect of the group supported the assumption of a different fundamental frequency (F_0) in individuals with and without ASD spectrum condition, $F(1, 76) = 15.51$, $p = 0.0002$, $\eta^2_G = 0.024$. As expected, there was an additional main effect of gender, reflecting the on average higher pitched voices of women, $F(1, 76) = 559.73$, $p < 0.0001$, $\eta^2_G = 0.86$.

Further analyzing HNR, a main effect of group was observed, $F(1, 76) = 7.97$, $p = 0.0061$, $\eta^2_G = 0.055$, pointing towards a higher HNR in individuals with ASD. As expected, there was an additional main effect of gender, $F(1, 76) = 60.78$, $p < 0.0001$, $\eta^2_G = 0.42$.

We found no significant evidence for group differences regarding jitter, shimmer, or energy (all $p > 0.01$, corrected for multiple comparisons).

For the machine-learning approach, we also included Mel-frequency cepstral coefficients (MFCCs) and were able to reach an AUC of 0.77. The exploratory in-depth analysis of the classification results showed no significant evidence for different accuracy of classifying male and female participants.

Clinical expert-based analysis

Given that misclassification of individuals with ASD by automated analysis of the SIT could be due to the paradigm not capturing enough information, we asked clinical experts in ASD diagnosis to rate the non-verbal behavior of participants from the videos to estimate the informative value of the SIT recordings. The clinical experts correctly recognized most of the HC individuals and more than half of the individuals with ASD. Among the eight psychologist/psychiatrist expert raters the accuracy ranged between 0.56 and 1 with a mean of 0.71. There was no evidence that experts classified more female than male individuals correctly ($X = 6.00$, $p = 0.11$). More experienced raters (months of diagnostic or therapeutic work with individuals with ASD) showed a higher percentage of correct classifications ($r_s = 0.72$, $p = 0.044$).

Comparison of ML-classifier and experts

Combining all features (gaze behavior, voice, and facial expressions), we reached an AUC of 0.78. Setting a threshold at 0.5 class probability, we reached an accuracy of 73%, which was according to McNemar's Test significantly better than a majority vote ($X = 9$, $p = 0.014$) and not significantly worse than the accuracy of the clinical experts ($p > 0.05$). The sensitivity was 67% and specificity 79%. The detection rate was slightly better for female (33 out of

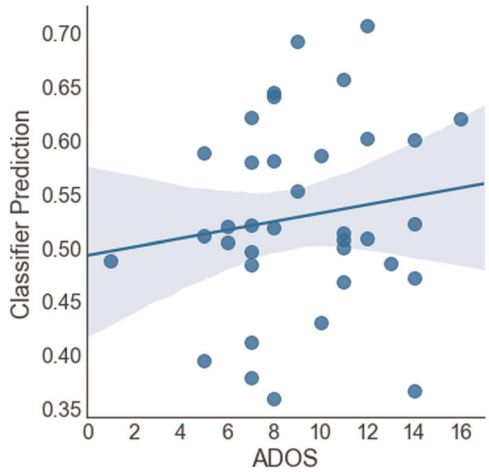

Fig. 2 Class probability for ASD and participant's ADOS score.

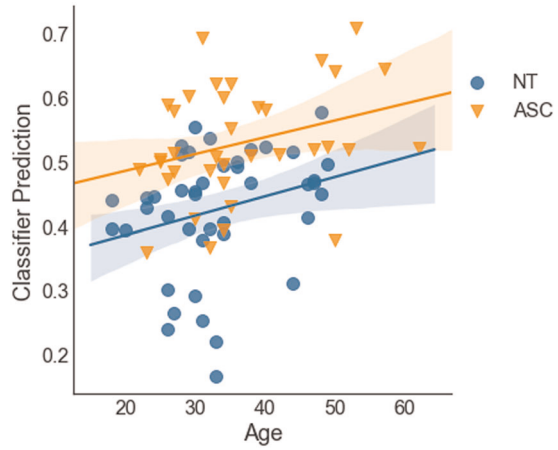

Fig. 3 Class probability for ASD and participant's age.

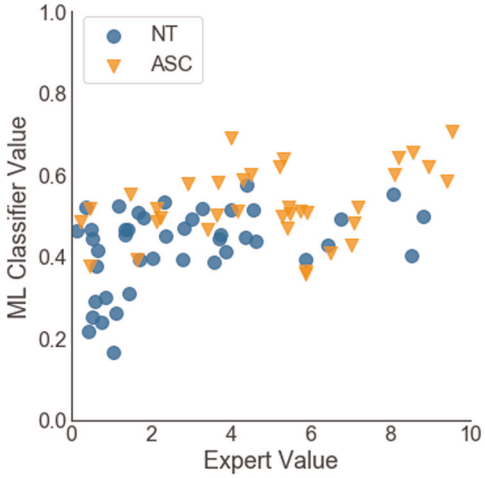

Fig. 4 Class probabilities based on machine-learning (ML) classification and expert ratings.

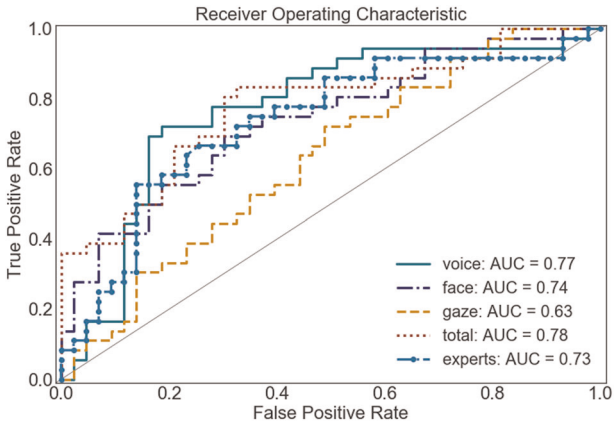

Fig. 5 ROC curves of all classifiers.

39) than male (25 out of 40) participants, $X = 3.88$, $p = 0.048$. Further, the class probability for ASD calculated by the classifier scaled positively with participants' ADOS score ($r_s = 0.48$, $p < 0.0001$, see Fig. 2) and with participant's age ($r_s = 0.35$, $p = 0.001$, see Fig. 3). For a comparative overview of the predicted class probabilities based on the machine-learning classification and expert ratings (on a scale of 0–10 and threshold at 5) see Fig. 4. For comparing the ROC curves of all classifiers see Fig. 5. Confusion matrixes of all classifiers are in Supplementary Tables 1–5.

DISCUSSION

We presented the simulated interaction task (SIT), a cost-efficient and time-efficient digital tool to identify social biomarkers, which we validated in an EMG study in healthy individuals and a clinical study using an automated digital approach in individuals with ASD. To the best of our knowledge, the SIT is the first fully standardized and computer-based measure of social interaction deficits. Using EMG as well as computer-vision-based analyses, we showed that the SIT reliably evoked positive and negative emotional expressions and as such represents a naturalistic paradigm for measuring non-verbal social behavior. Results of the ASD study suggest that the machine-learning-based automated analysis of facial expressions, gaze and voice holds potential to measure social phenotypical behavior and supplement traditional clinical assessment of social interaction deficits.

In the preparatory EMG study, participants smiled more when the actress talked about both liked and disliked food, compared to when she explained how she sets a table for dinner. Listening to the actress's disliked food, they also showed more facial expressions of disgust. Both social smiling and mimicking of joy and disgust replicates literature about typical human behavior in social interactions.[46–48] In line with the results of the EMG study, participants in the ASD study also expressed more positive facial expressions during the positive part of the conversation and tended to show more negative expressions during the negative part. Thus, the results of both studies validate the SIT as a tool to evoke naturalistic non-verbal social behavior (Table 3).

The preparatory study showed that across the whole conversation, individuals with more pronounced autistic traits expressed less smiling and more frowning. Based on these findings, we selected AUs representing smiling, frowning, and expressions of disgust for the ASD study to analyze their descriptive and diagnostic potential.

Using the SIT in a clinical sample of individuals with ASD, we showed group differences in the expression of facial emotions and voice modulation. Thus, the automated measurement of non-verbal social communication behavior holds potential to complement the phenomenological understanding of ASD. Individuals with ASD expressed less social smiling and less mimicry of positive facial expressions than HC individuals. This is in accordance with

Table 3. Mean standardized activity of each muscle in each conversation part.

	Zygomaticus major (z-stand.)	Corrugator supercilii (z-stand.)	Levator labii (z-stand.)
Dinner preparation (neutral)	−0.21 (SD: 0.86)	0.10 (SD: 1.21)	−0.14 (SD: 0.77)
Liked food (positive)	0.03 (SD:0.94)	−0.05 (SD: 0.83)	0.06 (SD:1.04)
Disliked food (negative)	0.15 (SD:1.22)	−0.04 (SD: 0.92)	0.06 (SD:1.12)

previous studies indicating reduced spontaneous mimicry in individuals with ASD.[25,49] Further, we are replicating the finding of a recent study using computer-based facial expression analysis that reported fewer happy expressions in individuals with ASD.[50]

Further, individuals with ASD spoke on average with higher pitch and higher HNR than NTs. Higher pitch has been described for individuals with ASD before,[51,52] although some studies reported null-results.[53] The naturalistic setting of the SIT resembling a common video chat situation might have enabled us to detect these differences more sensitively than previous studies. Interpretation of our HNR result is not straightforward, given the lack of reference studies. One study with children with ASD found a negative association between HNR median and disorder's severity.[54] In contrast, a study focusing on adolescents with ASD[55] found an association of higher median HNR and perceived voice awkwardness. Given this heterogeneity and given that HNR changes with age,[56] it is difficult to generalize those results to our study sample of adults with ASD. In addition, it cannot be ruled out that individuals with ASD differ from those without ASD regarding their ratios of speaking and pauses, which might have influenced the results. In general, quantitative evidence for voice differences in ASD is lacking, as a recent meta-analysis concluded.[53] Given the relatively high explanatory power for diagnosis in our study, however, further studies are needed to elucidate the role of HNR in ASD.

We found no significant evidence for group differences regarding gaze behavior, although many studies have reported differences before[57] and aberrant gaze has recently been discussed as a potential biomarker of ASD.[58] An explanation for our null-result might lie in the fact that the appearance-based gaze estimation used here, in comparison to eye tracking methods that were used previously, was not sensitive enough to detect the subtle gaze dysfunctions present in ASD. The results of the machine-learning-based analysis described below are in favor of this interpretation.

Comparing the groups regarding single features of socio-emotional behavior can be informative. However, the problem of multiple testing remains, as the automated analysis allows the comparison of many different aspects of behavior. Thus, we see a clear advantage of multi-dimensional approaches like machine learning that allow the capturing of the multivariate and integrated nature of naturalistic social behavior and analyze its diagnostic value.[59] Further, as we carefully split our data in train and test sets, we received information about the predictive value of the behavioral differences for individuals, which were not part of the training sample for the model.

The automated analysis of the SIT reached an accuracy of 73% at a set threshold of 0.5 class probability. The predicted class probabilities were associated positively with participants' overall ASD symptom severity. Only few studies using machine learning to detect ASD reached higher accuracies based on, e.g. automatic analysis of upper-limb movements,[34] or eye movements in a face-recognition task.[35] However, both studies used small sample sizes and were directed towards more affected children and not high-functioning adults with autism, which present with more subtle deficits. Further, those studies used high-cost apparatus, not suited for large-scale testing. In contrast, the application of the SIT only requires a standard PC, a camera, and a microphone. Thus, individuals can take the SIT at their home computers, which ensures, especially for individuals with autism, a more naturalistic setting than in a laboratory, leading to higher external validity and reducing confounding factors such as anxiety.

Based on facial expressions, more women than men could be classified correctly. One possible explanation might be that women with autism require more severe symptoms to receive an ASD diagnosis.[60] As our sample includes only individuals with ASD that were already diagnosed, the female individuals with ASD might have been easier to detect automatically based on their more dysfunctional facial expressions than the male participants.

Both classifications on all facial AUs available via OpenFace and on all voice characteristic including MFCC revealed good results pointing to the value of multidimensional representations of phenotypes. The classification based on gaze behavior was remarkably worse than the prediction based on vocal and facial behavior. Post hoc analysis of the gaze's variance suggests that this is likely due to the low precision of the eye gaze measurement. Also, the null finding of group differences in gaze behavior has to be interpreted with caution, as automated gaze tracking might not be accurate enough for computation of velocity and acceleration. Due to the relevance of gaze behavior in autism spectrum conditions,[31] further research should follow up on this with using state of the art eye-tracking devices to estimate the information value of gaze within a simulated interaction.

The classification based on the short recording of overall social communication behavior was not significantly worse than the clinical experts in detecting ASD diagnostic status. However, the judgment performance of the expert raters varied considerably from approximately chance level to 100% accuracy and scaled positively with experience. This can be seen as a strong argument for the application of the SIT, which is reliable as well as cost-efficient and time-efficient, while the clinical assessment does not only require intensive training but also, as our data indicates, continuous experience. Nevertheless, it should be highlighted that the SIT is not meant to substitute traditional clinical assessment but rather to allow additional information about the client's social communication abilities at an early stage of the diagnostic process.

Despite the substantial prevalence of ASD of 1 in 59 children,[61] there is only a limited amount of practitioners with expertise in diagnosing ASD.[62] As a result, many individuals, especially with high-functioning ASD, are diagnosed late[18] or not at all,[63] resulting in substantial burden.[64] In contrast to clinical interviews that require trained experts, the SIT offers the potential of a widely accessible and easily applicable screening tool for ASD, even in rural areas with limited access to clinical care. Therefore, the SIT may enable time-efficient detection of social interaction deficits and low-cost screenings of large populations. However, it is important to keep in mind that the SIT was designed as a tool to enrich and standardize the assessment of social communication behavior to supplement clinical diagnosis, screening, and treatment monitoring instead as a stand-alone diagnostic or screening tool. Nevertheless, it is crucial to understand the impact on individuals and health care systems of employing such automated behavior measurement tools and providing feedback based on them. On the one hand, the SIT offers a more standardized, more accessible and cost-effective measurement of non-verbal social communication behavior than clinical interviews and thus could improve screening, diagnostic, or monitoring processes. On the

other hand, especially screenings and unsupervised feedback might result in more individuals seeking care and lead to higher costs for the health system. It is warranted to carefully define the area of use and evaluate benefits, costs, and risks.

The importance of monitoring and characterizing social communication behavior is not limited to ASD, but applies to many other psychiatric conditions, as differences in recognizing and signaling emotions are common phenomena in mental disorders and vary with disorder severity, e.g., in psychosis,[4,65] depression,[5] bipolar disorder,[6] and substance use disorders.[7] Thus, the SIT may serve not only as a diagnostic instrument but also as a tool for measuring treatment-based outcomes in other conditions than ASD. Thus, we believe it is warranted to explore the SIT's specificity and sensitivity in larger clinical samples with different psychiatric conditions.

Interestingly, older participants in both groups were more often labeled as autistic than young participants by the machine-learning classifiers, as the classifier was probably picking up a tendency of higher age in the ASD group. This age bias points to the strong need of careful testing for confounders and biases in training data for all kind of machine-learning diagnostic methods.

Aiming for a high standardization, we prerecorded the actress' part of the conversation. Thus, she responds in exactly the same way to each participant at the same time point—independent of participant's response content or length. That being said, we took several measures to make the conversation flow as naturally as possible. First, the conversation is set up in such a way—and the participant is informed about this during the instructions—that the actress poses a series of questions, which the participant is asked to provide an answer for. Thus, the participant expects those questions rather than tailored replies to her answers. In addition, the participants' answer sections were made as natural as possible by the actress displaying empathic listening behavior, such as e.g. nodding, which would be appropriate for the kind of small talk conversation at hand. Second, we gave clear instructions (and provided an exercise) to answer in three to four sentences to the actress' questions: the conversation actually starts with a trial where the actress tells the participant how she got into the institute and asks the participant to describe how she got into the institute in three to four sentences.

However, future versions of the task would benefit from tailored answers of the actress that are adaptive to the length of the interactants' answers. For this first version of the SIT we aimed at the highest possible standardization though and asked the participants to behave as they would in a real interaction. In accordance, the majority of participants in the EMG-study reported post hoc that they behaved similar to a real conversation when talking to the actress. To further evaluate this aspect in the future, the current version of the SIT ends with three follow-up questions, how naturalistic the participant perceived the interaction. Further measures should be implemented to allow for home setting use (e.g. automated checking of camera and microphone) and research should be carried out to evaluate the accuracy of the SIT in a home setting, as self-administration of tests often leads to diminished accuracy.

Future studies should be undertaken to compare the behavior in the simulated interaction with behavior in a face-to-face control condition. A live setting with the same topic and a dialog script could inform about the validity of the simulated procedure. As similarity enhances facial mimicry,[66] further version of the SIT could provide different interaction partners matching the participant's gender, age, or ethnical background.

Taken together, combining standardization with a naturalistic paradigm, the SIT allows to objectively quantify social communication behavior and qualifies as a cost-efficient and time-efficient digital tool to detect social biomarkers in ASD.

METHOD

Preparatory study

In a facial EMG preparatory study (cf. Supplementary Information), we assessed facial movement behavior via the SIT in a sample of healthy male individuals. Making use of the high precision of EMG, we aimed at the precise description of non-verbal facial behavior in the simulated dialog to select relevant features for the main ASD study.

Participants

Thirty-seven adults with ASD (18 females, mean age = 36.89 years, range = 22–62) and 43 HC individuals (22 females; mean age = 33.14 years, range = 18–49) were enrolled in the study. This sample size exceeded the required minimum sample size of 78 estimated by a power analysis (evaluated for independent *t*-tests with power = 0.80, $\alpha = 0.05$ and a moderate effect size $\rho = 0.40$, based on meta-analysis of vocal characteristics[53] and facial expressions[67] as marker for autism). Individuals with ASD were recruited via the ASD outpatient clinic of the Charité— Universitätsmedizin Berlin. All of the participants with ASD had previously received a diagnosis of Asperger syndrome, atypical ASD or childhood ASD based on ICD-10 criteria.[68] The diagnostic procedure included the ADOS ($n = 35$[69]) and, for patients with available parental informants, the ADIR ($n = 21$[70]).

Only participants without current antipsychotic and anticonvulsant medication as well as comorbid neurological disorders were included. In addition, we set the maximum age for inclusion at 65 years to avoid possible age-related neurodegeneration. Furthermore, we used a German vocabulary test for verbal intelligence (Wortschatztest)[71] to ensure sufficient language command. In the control group, any history of psychiatric disorder led to exclusion. Based on these criteria, we excluded three participants.

Procedure

The study took place in a quiet laboratory with constant lighting conditions. After having given informed consent, participants were asked to sit in front of a computer screen. The participants were informed that they were to have a short conversation with a woman, whose part was recorded before, and were encouraged to behave as they would in a natural conversation. The experimenter then left the room and the participant started the SIT. During the whole simulated interaction, the participant's face, gaze behavior, and voice were video and audio recorded, respectively, and analyzed later using computer-based technologies. The video of the participant was recorded automatically and timestamped to later align the actress's and participant's recordings. Participant's autistic traits were assessed via the Autism-Spectrum Quotient.[72] All participants gave written informed consent prior to participation. All individuals displayed in figures have given their consent for publication. The study was approved by the ethics committee of the Charité—Universitätsmedizin Berlin and was conducted in accordance with the Declaration of Helsinki.

Simulated interaction task

The SIT (Fig. 6) is a simulated social interaction designed as a conversation between the participant and an actress about food preferences and dinner preparation, where the part of the actress is prerecorded. In the first part of the conversation (3.08 min), the actress introduces herself and asks the participants for their names for a short warm-up. Next, she explains the task and the topics of the following conversation to the participants and, as an example, asks the participant to describe how they got to the institute. This part of the conversation represents a warm-up for the dialog and is not analyzed. Thereafter, the actress introduces the three topics that the conversation will be about, i.e., dinner preparations and liked/disliked food and announces that participants will be given time to think about their answers before the participant starts the task by button press. This preparatory phase ensures the conversation's flow and thus approximates real-life conversation conditions.

In the following recording (3.2 min), the actress addresses the participant in three excerpts about (i) how she usually sets a table for dinner (26 s), (ii) food she likes (24 s), and (iii) food she dislikes (26 s). Following each section, she asks the participant about the same respective preferences and information. After each question, the participant has about half a minute time to answer (see Fig. 7). While the participant answers, the actress displays empathic listening behaviors, e.g., she smiles

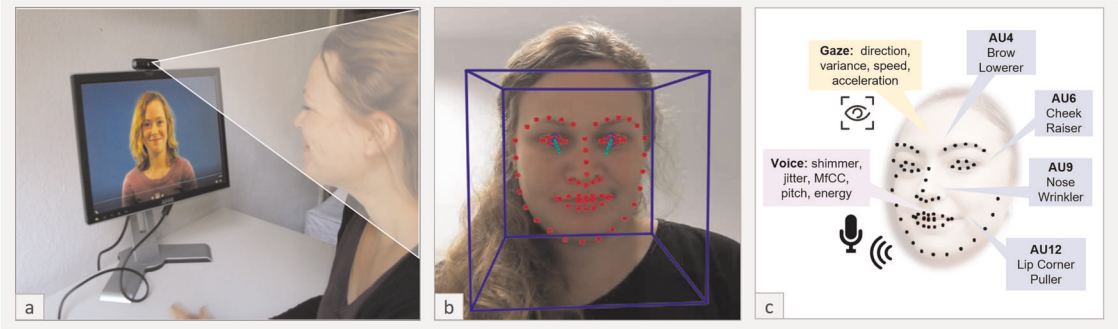

Fig. 6 Experimental setting und automated analysis of SIT. a Neurotypical participant taking the SIT. **b** Face-tracking and gaze-tracking using OpenFace. **c** Facial landmarks and features of main interest; written informed consent was obtained from the persons to have their photos used in this study.

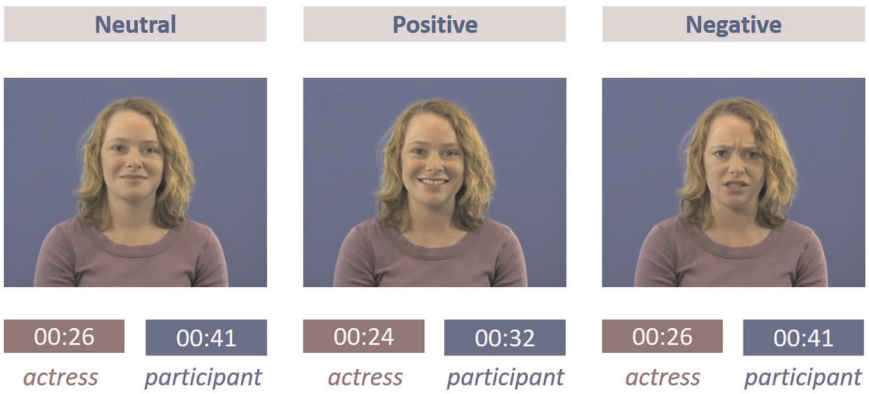

Fig. 7 Timing for each excerpt of SIT. Written informed consent was obtained from the person to have her photos used in this study.

and nods at the participant. Thus, the SIT allows to objectively measure qualitative and quantitative differences in social communication behavior with a high-level of standardization and in an interactive naturalistic manner.

Recent research in the area of social cognition has emphasized the strong need for more interactive experimental social tasks for basic research and the clinical diagnosis process.[73] However, interactive paradigms are often more time consuming and financially costly, as they require a second participant, a confederate or professional to interact with the participant. Standardization is furthermore a challenge as the participant's counterpart rarely interacts in the exact same manner. In contrast, the SIT uses a video conversation setting, which today represents a familiar interaction setting for many individuals, as they are common both in private and business context. Using a prerecorded partner, the SIT does not demand a second person for testing. We chose food preferences as conversation topic for the SIT, as this represents a typical small talk subject that allows for the elicitation of positive and negative emotions in a short amount of time without involving highly personal or sensitive experiences.[74] Moreover and importantly, conversation excerpts vary considerably less in length than would be the case with other topics that have been used in experimental settings (e.g., favorite films) which thus facilitates standardization. Given those considerations, the SIT represents a quasi-naturalistic setting, balancing free interaction and standardization.

During the shooting of the video, in order to generate natural dialog behavior, the actress actually conversed with the film director about food preferences, which thus not only involved her talking about her preferences but also listening to his. The actress was instructed to show empathic communication behavior, entailing generally positive attitude towards the interaction partner, including confirmative nodding und social smiling. The duration of these parts for the SIT was estimated based on time needed from sample participants.

The SIT allows a non-intrusive measurement of non-verbal communication behavior, which demands a computer with camera and microphone only and can thus be conducted in a laboratory and home setting. In particular clinical studies may benefit from the non-intrusive nature of the task, as additional equipment such as EMG electrodes might hinder naturalistic behavior and introduce other confounds (e.g. aversive reactions in touch-sensitive ASD patients).

We validated and quantified the actress' non-verbal behavior by analyzing her video with automatic facial expression analysis, focusing on AUs most relevant for social behavior and the specific emotions addressed (AU6, AU12, AU9, AU4). The actress shows more expression of disgust (AU4: $M = 0.23$, AU9: $M = 0.07$) and less expression of joy (AU6: $M = 0.61$, AU12: $M = 0.72$) when she speaks about disgusting food versus speaking about her favorite food (AU4: $M = 0.06$, AU9: $M = 0.05$, AU6: $M = 0.97$, AU12: $M = 1.2$).

The same pattern, indicating mimicry behavior, is evident when she listens to the participant talking about their favorite (AU4: $M = 0.02$, AU9: $M = 0.04$, AU6: $M = 1.15$, AU12: $M = 1.36$,) versus non-favorite food (AU4: $M = 0.13$, AU9: $M = 0.06$, AU6: $M = 0.95$, AU12: $M = 1.11$).

As expected for empathic listening and representing affiliative behavior, a comparison of the actress listening versus talking revealed that the actress smiles more when she listens (AU6: $M = 0.53$; AU12: $M = 0.92$) than when she speaks (AU6: $M = 0.33$, AU12: $M = 0.79$).

Automatic data analysis

We analyzed face, gaze, and voice recordings using the open source toolkit OpenFace[42] and python library Librosa[75] and Parselmouth Praat Scripts in Python by David Feinberg.[76] If not otherwise specifiecd, all statistical tests are two-sided. Based on the literature and the EMG preparatory study, we focused on features representing typical social and affiliative behavior and explain them in more detail in the following sections:

OpenFace identifies small observable movement in facial AUs using computer-based analysis. We followed the same data analysis procedure for this automatic approach as for the EMG data. First, we excluded every participant who was not tracked successfully in more than 90% of the frames or with a mean confidence below 0.75. The frame confidence is

computed by OpenFace via training a separate confidence network that is trained to predict the expected landmark detection error. Second, we measured whether facial expressions matched the predefined emotion for each conversation part (e.g., happiness in the excerpts dealing with positive food preferences). To this end the intensity of the corresponding emotional facial expression was measured, operationalized as the mean activity of the relevant AUs based on Ekman and Friesen,[77] i.e., for joy AU6 (Orbicularis Occuli) and AU12 (Zygomaticus Major) and for disgust AU9 (Levator Labbi) and AU4 (Corrugator Supercilli). The neutral part was used as a baseline measurement for the other two parts.

Further, we assessed the participant's mimicking of the actress' facial expressions: For each relevant AU, we calculated a correlation between the actress' and the participant's activity within 10-s timeframes. Tracking affiliative behavior and based on the finding of the EMG preparatory study, we further measured the individual's social smiling during the entire conversation (AU12 and AU6).

Before analyzing the gaze behavior, we excluded frames tracked unsuccessfully or with confidence lower than 0.75 (on a scale of 0–1), as well as participants that were not tracked successfully (with <90% of all frames recognized).

OpenFace provides gaze angle values for each frame, which are provided in *x* and *y* coordinates with the camera as the reference point. To account for participants' height, all values were centered around the median. We calculated the mean of the absolute values (this equals the mean deviation of the median of the gaze), and the means for the first derivate (speed of the gaze) and the second derivate (acceleration of gaze) for the whole conversation. Further, we measured the participant's mimicking of the actress' facial gaze using correlations within 10 s timeframes.

We analyzed the audio recording of each participant's voice using the python library Librosa[75] and Parselmouth Praat Scripts in Python by David Feinberg[76] to extract the following prosodic features for each frame: *f*0 (fundamental frequency of vocal oscillation), jitter (pitch perturbations), and shimmer (amplitude perturbations) and the root-mean-square energy. All of these features are frequently used in voice analysis and have shown to vary in different clinical conditions.[78,79] To compare the two groups, we calculated the respective means over all frames of the whole conversation for each feature for each subject. For the machine-learning analysis of the diagnostic value of the voice, we additionally extracted with Librosa first 40th Mel-frequency cepstral coefficients, which capture essential information from voice signal and have been used in automatic speech recognition as well as in voice disorder classification.[80]

Machine-learning analysis

This work builds on and extends a preliminary conference presentation directed at a machine-learning audience.[41] In addition to the results of Drimalla et al.,[41] this paper compares the machine-learning approach to ratings of clinical experts, explores differences in facial features, voice features, and gaze features between subject groups. To investigate the predictive value of the features from each of the three domains (facial expression, gaze, and voice), we used machine-learning methods. For machine-learning analysis, the library scikit-learn (https://scikit-learn.org) for Python was used.

First, to gain a better representation for the machine-learning analysis, we extracted secondary features. For facial AUs and gaze angles, we calculated mean, standard deviation, minima, maxima, time points of maximum, skewness, and kurtosis. The calculations were done separately for each of the seven parts of the conversation. As a result, we used 715 features for the facial expression and 85 features for the gaze behavior. For voice analysis, we calculated the first forty mel frequency cepstral coefficients (MFCC), energy, the mean and standard deviation of the fundamental frequency and different measures of shimmer and jitter. Due to computational power, we limited us to these 58 primary features. Gender was included as a feature in all models to ensure that the model could select certain features as being more predictive for certain genders.

For each participant, we built only one feature vector that contains the aggregate statistical measures across all video segments. Thus, the data for the machine-learning analysis did not include any nested repeated measures. Using this feature representation, we conducted a machine-learning analysis to explore the diagnostic value of non-verbal social behaviors (gaze/face/voice), i.e., the prediction of whether a participant had received a diagnosis of ASD. We built models to map the feature vectors to a value of a binary (ASD vs. NT) decision function involving balanced classes.

We compared different machine-learning approaches to solve this problem in a previous paper.[41] In this paper, we used a random forest approach,[81] which classifies the participants averaging the results of an ensemble of 1000 different decision trees on different subsets of data and input variables. The maximum depth of the trees and the minimum number of samples per leaf were the hyperparameters: For depth of trees the used values were [1, 2, 4, 8, 16, 32, 64], and for samples per leaf [1, 2, 4, 8, 16, 32, 64].

To train and test the machine-learning models, we used a leave-one-out cross-validation. To calculate the AUC, we trained separately for each participant a model with the data of the other participants and then tested each model only on the respective left-out participant. No data of the test-participant was used for training the model. Inside the training data, we tuned the hyperparameter with a nested three-fold-cross-validation.

Analysis of the data by clinical experts

Eight psychologists/psychiatrists with ample experience in diagnosing and treating individuals with ASD evaluated 10 randomly selected videos of different participants each, which were unknown to them. To match the rating task as closely as possible to the machine-learning task, experts were asked to focus on non-verbal behavior only (facial expressions, voice characteristics, and gaze behavior). After watching a video, they were instructed to rate on a visual analog scale where they would locate the participant on the autism spectrum, i.e., to estimate the amount of autistic traits/symptomatology that the individual presents with. The scale ranged from "neurotypical" to "ASD" with a vertical line in the middle that indicated the point at which autism symptomatology reaches clinical significance and warrants a diagnosis of ASD. We assigned the videos randomly to the raters with no predefined proportion of individuals with and without ASD. Each video was rated once.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The datasets generated and analyzed during the current study are not publicly available due to privacy restrictions of the video data but are available from the corresponding author on reasonable request.

CODE AVAILABILITY

The SIT is available from the corresponding author on reasonable request. From 2024 on, the SIT is publicly available online via https://edoc.hu-berlin.de/handle/18452/21028 under CC BY-NC-ND 4.0 license. The code used for analyzing the data is publicly available on Github: https://github.com/drimalla/ml-autism.

Received: 11 June 2019; Accepted: 17 January 2020;
Published online: 28 February 2020

REFERENCES

1. Cotter, J. et al. Social cognitive dysfunction as a clinical marker: a systematic review of meta-analyses across 30 clinical conditions. *Neurosci. Biobehav. Rev.* **84**, 92–99 (2018).
2. Lozier, L. M., Vanmeter, J. W. & Marsh, A. A. Impairments in facial affect recognition associated with autism spectrum disorders: a meta-analysis. *Dev. Psychopathol.* **26**, 933–945 (2014).
3. Uljarevic, M. & Hamilton, A. Recognition of emotions in autism: a formal meta-analysis. *J. Autism Dev. Disord.* **43**, 1517–1526 (2013).
4. Barkl, S. J., Lah, S., Harris, A. W. F. & Williams, L. M. Facial emotion identification in early-onset and first-episode psychosis: a systematic review with meta-analysis. *Schizophrenia Res.* **159**, 62–69 (2014).
5. Kupferberg, A., Bicks, L. & Hasler, G. Social functioning in major depressive disorder. *Neurosci. Biobehav. Rev.* **69**, 313–332 (2016).
6. Bertsch, K., Hillmann, K. & Herpertz, S. C. Behavioral and neurobiological correlates of disturbed emotion processing in borderline personality disorder. *Psychopathology* **51**, 76–82 (2018).
7. Castellano, F. et al. Facial emotion recognition in alcohol and substance use disorders: a meta-analysis. *Neurosci. Biobehav. Rev.* **59**, 147–154 (2015).

8. Thorup, E., Nyström, P., Gredebäck, G., Bölte, S. & Falck-Ytter, T. Altered gaze following during live interaction in infants at risk for autism: an eye tracking study. *Mol. Autism* **7**, 12 (2016).
9. Davies, H. et al. Facial expression to emotional stimuli in non-psychotic disorders: a systematic review and meta-analysis. *Neurosci. Biobehav. Rev.* **64**, 252–271 (2016).
10. Nasreddine, Z. S. et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatrics Soc.* **53**, 695–699 (2005).
11. Folstein, M. F., Folstein, S. E. & McHugh, P. R. Mini-mental state. A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* **12**, 189–198 (1975).
12. Conzelmann, K., Weis, S. & Süß, H.-M. New Findings About Social Intelligence. *J. Individ. Dif.* **34**, 19–137 (2013).
13. Lord, C. et al. The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J. Autism Dev. Disord.* **30**, 205–223 (2000).
14. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, DSM-5*. (American Psychiatric Association, Arlington, VA, 2013).
15. Lord, C., Rutter, M. & Le Couteur, A. Autism Diagnostic Interview-Revised. A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J. Autism Dev. Disord.* **24**, 659–685 (1994).
16. Fusar-Poli, L. et al. Diagnosing ASD in adults without ID: accuracy of the ADOS-2 and the ADI-R. *J. Autism Dev. Disord.* **47**, 3370–3379 (2017).
17. Bastiaansen, J. A. et al. Diagnosing autism spectrum disorders in adults: the use of autism diagnostic observation schedule (ADOS) module 4. *J. Autism Dev. Disord.* **41**, 1256–1266 (2011).
18. Barnard, J., Harvey, V. & Potter, D. *Ignored or Ineligible? The Reality for Adults with Autism Spectrum Disorders* (National Autistic Society, 2001).
19. Harms, M. B., Martin, A. & Wallace, G. L. Facial emotion recognition in autism spectrum disorders. A review of behavioral and neuroimaging studies. *Neuropsychol. Rev.* **20**, 290–322 (2010).
20. Moore, E., Clements, M. A., Peifer, J. W. & Weisser, L. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE Trans. Bio-Med. Eng.* **55**, 96–107 (2008).
21. Cohn, J. F. et al. In *Proc. 2009 3rd International Conference on Affective Computing and Intelligent Interaction* (ed. Staff, I.) 1–7 (IEEE, 2009).
22. Alghowinem, S. et al. Cross-cultural detection of depression from nonverbal behaviour. In *11th IEEE International Conference and workshops on automatic face and gesture recognition (FG)*. Vol. 1 (IEEE, 2015).
23. Sigman, M. D., Kasari, C., Kwon, J.-H. & Yirmiya, N. Responses to the negative emotions of others by autistic, mentally retarded, and normal children. *Child Dev.* **63**, 796 (1992).
24. Reddy, V., Williams, E. & Vaughan, A. Sharing humour and laughter in autism and Down's syndrome. *Br. J. Psychol.* **93**, 219–242 (2002).
25. McIntosh, D. N., Reichmann-Decker, A., Winkielman, P. & Wilbarger, J. L. When the social mirror breaks: deficits in automatic, but not voluntary, mimicry of emotional facial expressions in autism. *Dev. Sci.* **9**, 295–302 (2006).
26. Yoshimura, S., Sato, W., Uono, S. & Toichi, M. Impaired overt facial mimicry in response to dynamic facial expressions in high-functioning autism spectrum disorders. *J. Autism Dev. Disord.* **45**, 1318–1328 (2015).
27. Zhao, S., Uono, S., Yoshimura, S., Kubota, Y. & Toichi, M. Atypical gaze cueing pattern in a complex environment in individuals with ASD. *J. Autism Dev. Disord.* **47**, 1978–1986 (2017).
28. Wieckowski, A. T. & White, S. W. Eye-gaze analysis of facial emotion recognition and expression in adolescents with ASD. *J. Clin. Child Adolesc. Psychol. 53* **46**, 110–124 (2017).
29. Tanaka, J. W. & Sung, A. The "Eye Avoidance" hypothesis of autism face processing. *J. Autism Dev. Disord.* **46**, 1538–1552 (2016).
30. Madipakkam, A. R., Rothkirch, M., Dziobek, I. & Sterzer, P. Unconscious avoidance of eye contact in autism spectrum disorder. *Sci. Rep.* **7**, 13378 (2017).
31. Kirchner, J. C., Hatri, A., Heekeren, H. R. & Dziobek, I. Autistic symptomatology, face processing abilities, and eye fixation patterns. *J. Autism Dev. Disord.* **41**, 158–167 (2011).
32. Nakai, Y., Takiguchi, T., Matsui, G., Yamaoka, N. & Takada, S. Detecting abnormal word utterances in children with autism spectrum disorders: machine-learning-based voice analysis versus speech therapists. *Percept. Mot. Skills* **124**, 961–973 (2017).
33. Nasir, M., Jati, A., Shivakumar, P. G., Nallan Chakravarthula, S. & Georgiou, P. in *Proc. 6th International Workshop on Audio/Visual Emotion Challenge—AVEC '16* (eds. Valstar, M. et al.) 43–50 (ACM Press, New York, NY, USA, 2016).
34. Crippa, A. et al. Use of machine learning to identify children with autism and their motor abnormalities. *J. Autism Dev. Disord.* **45**, 2146–2156 (2015).
35. Liu, W., Li, M. & Yi, L. Identifying children with autism spectrum disorder based on their face processing abnormality. A machine learning framework. *Autism Res.* **9**, 888–898 (2016).
36. Gliga, T., Bedford, R., Charman, T. & Johnson, M. H. Enhanced visual search in infancy predicts emerging autism symptoms. *Curr. Biol.* **25**, 1727–1730 (2015).
37. Hashemi, J. et al. A Computer Vision Approach for the Assessment of autism-related behavioral markers. In *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*. 1–7 (IEEE, 2012).
38. Egger, H. L. et al. Automatic emotion and attention analysis of young children at home: a ResearchKit autism feasibility study. *npj Digital Med.* **1**, 1 (2018).
39. Lipinski, S., Blanke, E. S., Suenkel, U. & Dziobek, I. Outpatient psychotherapy for adults with high-functioning autism spectrum condition: utilization, treatment satisfaction, and preferred modifications. *J. Autism Dev. Disord.* **49**, 1154–1168 (2019).
40. Vogeley, K., Kirchner, J. C., Gawronski, A., van Tebartz Elst, L. & Dziobek, I. Toward the development of a supported employment program for individuals with high-functioning autism in Germany. *Eur. Arch. Psychiatry Clin. Neurosci.* **263**, S197–S203 (2013).
41. Drimalla, H. et al. *in Machine Learning and Knowledge Discovery in Databases* (eds. Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N. & Ifrim, G.) 193–208 (Springer International Publishing, Cham, 2019).
42. Baltrusaitis, T., Zadeh, A., Lim, Y. C. & Morency, L.-P. Openface 2.0: Facial behavior analysis toolkit. In *13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, 59–66 (IEEE, 2018).
43. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874 (2006).
44. Sussman, J. E. & Sapienza, C. Articulatory, developmental, and gender effects on measures of fundamental frequency and jitter. *J. Voice* **8**, 145–156 (1994).
45. Zäske, R., Skuk, V. G., Kaufmann, J. M. & Schweinberger, S. R. Perceiving vocal age and gender: an adaptation approach. *Acta Psychol.* **144**, 583–593 (2013).
46. Hess, U. & Bourgeois, P. You smile–I smile: emotion expression in social interaction. *Biol. Psychol.* **84**, 514–520 (2010).
47. Rymarczyk, K., Żurawski, Ł., Jankowiak-Siuda, K. & Szatkowska, I. Emotional empathy and facial mimicry for static and dynamic facial expressions of fear and disgust. *Front. Psychol.* **7**, 1853 (2016).
48. Hess, U. & Fischer, A. Emotional mimicry as social regulation. *Personal. Soc. Psychol. Rev.* **17**, 142–157 (2013).
49. Oberman, L. M., Winkielman, P. & Ramachandran, V. S. Slow echo: facial EMG evidence for the delay of spontaneous, but not voluntary, emotional mimicry in children with autism spectrum disorders. *Dev. Sci.* **12**, 510–520 (2009).
50. Owada, K. et al. Computer-analyzed facial expression as a surrogate marker for autism spectrum social core symptoms. *PLoS ONE* **13**, e0190442 (2018).
51. Filipe, M. G., Frota, S., Castro, S. L. & Vicente, S. G. Atypical prosody in Asperger syndrome: perceptual and acoustic measurements. *J. Autism Dev. Disord.* **44**, 1972–1981 (2014).
52. Sharda, M. et al. Sounds of melody–pitch patterns of speech in autism. *Neurosci. Lett.* **478**, 42–45 (2010).
53. Fusaroli, R., Lambrechts, A., Bang, D., Bowler, D. M. & Gaigg, S. B. Is voice a marker for Autism spectrum disorder? A systematic review and meta-analysis. *Autism Res.* **10**, 384–407 (2017).
54. Bone, D. et al. The psychologist as an interlocutor in autism spectrum disorder assessment: insights from a study of spontaneous prosody. *J. Speech Lang. Hearing Res.* **57**, 1162–1177 (2014).
55. Bone, D., Black, M. P., Ramakrishna, A., Grossman, R. B. & Narayanan, S. S. Acoustic-prosodic correlates of 'awkward' prosody in story retellings from adolescents with autism, In *INTERSPEECH-2015*, 1616–1620 (2015).
56. Ferrand, C. T. Harmonics-to-noise ratio. *J. Voice* **16**, 480–487 (2002).
57. Papagiannopoulou, E. A., Chitty, K. M., Hermens, D. F., Hickie, I. B. & Lagopoulos, J. A systematic review and meta-analysis of eye-tracking studies in children with autism spectrum disorders. *Soc. Neurosci.* **9**, 610–632 (2014).
58. Freedman, E. G. & Foxe, J. J. Eye movements, sensorimotor adaptation and cerebellar-dependent learning in autism: toward potential biomarkers and subphenotypes. *Eur. J. Neurosci.* **47**, 549–555 (2018).
59. Bzdok, D. & Meyer-Lindenberg, A. Machine learning for precision psychiatry: opportunities and challenges. *Biol. Psychiatry* **3**, 223–230 (2018).
60. Russell, G., Ford, T., Steer, C. & Golding, J. Identification of children with the same level of impairment as children on the autistic spectrum, and analysis of their service use. *J. Child Psychol. Psychiatry Allied Discip.* **51**, 643–651 (2010).
61. Baio, J. et al. Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2014. *Morbidity Mortal. Wkly. Rep. Surveill. Summaries (Wash., D. C.: 2002)* **67**, 1–23 (2018).
62. Crane, L. et al. Autism diagnosis in the United Kingdom: perspectives of autistic adults, parents and professionals. *J. Autism Dev. Disord.* **48**, 3761–3772 (2018).

63. White, S. W., Ollendick, T. H. & Bray, B. C. College students on the autism spectrum: prevalence and associated problems. *Autism: Int. J. Res. Pract.* **15**, 683–701 (2011).
64. Lehnhardt, F.-G. et al. Das psychosoziale Funktionsniveau spätdiagnostizierter Patienten mit Autismus-Spektrum-Störungen–eine retrospektive Untersuchung im Erwachsenenalter. *Fortschr. Neurol.-Psychiatr.* **80**, 88–97 (2012).
65. Kohler, C. G. et al. Static posed and evoked facial expressions of emotions in schizophrenia. *Schizophrenia Res.* **105**, 49–60 (2008).
66. Seibt, B., Mühlberger, A., Likowski, K. & Weyers, P. Facial mimicry in its social setting. *Front. Psychol.* **6**, 1122 (2015).
67. Trevisan, D. A., Hoskyn, M. & Birmingham, E. Facial expression production in autism: a meta-analysis. *Autism Res.* **11**, 1586–1601 (2018).
68. World Health Organization. *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research* (World Health Organization, 1993).
69. Lord, C. et al. The autism diagnostic observation schedule—generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J. Autism Dev. Disord.* **30**, 205–223 (2000).
70. Rutter, M., Le Couteur, A. & Lord, C. *Autism Diagnostic Interview-Revised*, Vol. 29, 30 (Western Psychological Services, Los Angeles, CA, 2003).
71. Schmidt, K.-H. & Metzler, P. *Wortschatztest (WST)*. (Beltz Test, Weinheim, 1992).
72. Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J. & Clubley, E. The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians. *J. Autism Dev. Disord.* **31**, 5–17 (2001).
73. Dziobek, I. Comment: towards a more ecologically valid assessment of empathy. *Emot. Rev.* **4**, 18–19 (2012).
74. Blackburn, K. G., Yilmaz, G. & Boyd, R. L. Food for thought: exploring how people think and talk about food online. *Appetite* **123**, 390–401 (2018).
75. Brian McFee et al. *librosa/librosa: 0.6.3* (Zenodo, 2019).
76. Feinberg, David R. *Parselmouth Praat Scripts in Python*. https://doi.org/10.17605/OSF.IO/6DWR3 (2019).
77. Ekman, P. & Friesen, W. V. *Facial Action Coding System*. (Consulting Psychologists Press, Palo Alto, CA, 1978).
78. Orozco-Arroyave, J. R. et al. Automatic detection of Parkinson's disease in running speech spoken in three different languages. *J. Acoust. Soc. Am.* **139**, 481–500 (2016).
79. Zhang, J., Pan, Z., Gui, C., Zhu, J. & Cui, D. Clinical investigation of speech signal features among patients with schizophrenia. *Shanghai Arch. Psychiatry* **28**, 95–102 (2016).
80. Eskidere, Ö. & Gürhanlı, A. Voice disorder classification based on multitaper mel frequency cepstral coefficients features. *Comput. Math. Methods Med.* **2015**, 956249 (2015).
81. Breiman, L. Random forest. *Mach. Learn.* **45**, 5–32 (2001).

ACKNOWLEDGEMENTS

H.D. was partially funded by the Berlin School of Mind and Brain under grant GSC 86/1-3. N.L. was partially funded by the German Research Foundation under grant LA/3270/1-1. We thank Christian Knauth for implementing the SIT into a stand-alone application for Windows and Mac. We acknowledge support by the German Research Foundation (DFG) and the Open Access Publication Fund of Humboldt-Universität zu Berlin.

AUTHOR CONTRIBUTIONS

H.D. designed the study concept, the experimental procedure and the stimuli, analyzed and interpreted the data, and wrote the manuscript. T.S. and N.L. oversaw and assisted with all the aspects of the machine-learning analysis. I.B. collected the data and made a contribution to the manuscript within the scope of her master thesis in Psychology. S.R. contributed to the conception of the study and the data acquisition. B.B. contributed to the conception of the study and oversaw the data acquisition. I.D. oversaw and assisted with all the aspects of the study design, data analysis, and writing process. All authors read and approved the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41746-020-0227-5.

Correspondence and requests for materials should be addressed to H.D.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.